Neuron
30 Perspective
CellPress

Working Memory 2.0

Earl K. Miller,[1,*] Mikael Lundqvist,[1] and André M. Bastos[1]
[1]The Picower Institute for Learning and Memory, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*Correspondence: ekmiller@mit.edu
https://doi.org/10.1016/j.neuron.2018.09.023

Working memory is the fundamental function by which we break free from reflexive input-output reactions to gain control over our own thoughts. It has two types of mechanisms: online maintenance of information and its volitional or executive control. Classic models proposed persistent spiking for maintenance but have not explicitly addressed executive control. We review recent theoretical and empirical studies that suggest updates and additions to the classic model. Synaptic weight changes between sparse bursts of spiking strengthen working memory maintenance. Executive control acts via interplay between network oscillations in gamma (30–100 Hz) in superficial cortical layers (layers 2 and 3) and alpha and beta (10–30 Hz) in deep cortical layers (layers 5 and 6). Deep-layer alpha and beta are associated with top-down information and inhibition. It regulates the flow of bottom-up sensory information associated with superficial layer gamma. We propose that interactions between different rhythms in distinct cortical layers underlie working memory maintenance and its volitional control.

Introduction

Working memory is the "sketchpad of conscious thought." It is the platform where we hold and manipulate thoughts and is foundational to the organization of goal-directed behavior (Chatham and Badre, 2015; Engle et al., 1999; Fuster, 1999; Goldman-Rakic, 1995; Just and Carpenter, 1992; Miller and Cohen, 2001; Vogel and Machizawa, 2004).

Starting with work by Fuster, Goldman-Rakic, and others, a wealth of data have shown that neurons in higher-order cortex, including the prefrontal cortex (PFC), show "delay activity"—elevated levels of spiking during memory delays of working memory tasks (Funahashi et al., 1989; Fuster and Alexander, 1971). For example, a stimulus is shown that must be remembered over a brief (one second or more) delay. The stimulus causes increased spiking. After it is gone, neurons continue to spike, typically at a lower rate but still above baseline levels (i.e., just before the stimulus). Everything we know suggests delay activity spiking helps maintain the working memory of the stimulus. We now also know that working memory involves much of the cortex. It engages executive functions associated with frontal cortex as well as posterior cortical areas that help maintain specific content (Fuster, 2015; Lara and Wallis, 2015; Miller and Cohen, 2001).

But how, exactly, does spiking do that? Under the "classic" model, delay activity reflects persistent spiking that keeps neural ensembles "online" in a continual state of activation. However, it is important to keep in mind that virtually all of the evidence for persistent spiking is based on the time-honored practice of averaging spiking over time and across trials. This was a necessity for performing statistical analyses, especially if the data were collected one neuron at a time (as it often was prior to the advent of multi-electrode recording). But this averaging can make spiking appear persistent, even though, in real time, e.g., on single trials, it is sparse (Lundqvist et al., 2016, 2018a; Shafi et al., 2007).

And there are issues with persistent spiking. Spikes are metabolically expensive. Memories held by persistent spiking alone can be labile because they are lost when activity is disrupted. Multiple items can be simultaneously held if each item engages non-overlapping ensembles (Almeida et al., 2015; Edin et al., 2009). But neural ensembles often have a high degree of overlap (Fusi et al., 2016; Rigotti et al., 2013; Warden and Miller, 2010). Plus, neurons optimize information when they spike sparsely and in bursts, not persistently (Lisman, 1997; Naud and Sprekeler, 2018). In other words, in the constant chatter of the brain, a brief scream is heard better than a constant whisper. Sparse spiking also allows multiple items to be multiplexed in time, preventing them from interfering with one another and simplifying the readout of working memory (different ensembles shout in turn instead of mumbling on top of each other; Bahramisharif et al., 2017; Lisman and Idiart, 1995; Lundqvist et al., 2011; Sandberg et al., 2003; Siegel et al., 2009). In fact, even sustained attention is not truly sustained. The brain samples the environment periodically (Buschman and Miller, 2010; Fiebelkorn et al., 2018; Helfrich et al., 2018; Landau and Fries, 2012; Schroeder et al., 2010; VanRullen, 2016). All this suggests that working memory (and cognition in general) is more complex than a simple persistence of spiking and average spike rates.

Further, a critical aspect of working memory has not enjoyed as much experimental effort as its maintenance functions. Volitional control is what makes working memory special. It is the fundamental function by which our brain wrests control of behavior from the environment and turns it to our own internal goals (Goldman-Rakic, 1995). We can choose what to think about and when and whether to act. Breakdown in volition is associated with psychiatric disease, like schizophrenia (Uhlhaas and Singer, 2010). Volition is, necessarily, a network phenomenon and thus not well addressed at the single-neuron level. Network properties can be examined with multiple-electrode recordings of multiple neurons and at the level of local field

potential (LFP) level, the summed activity of many neurons. During working memory tasks, there are LFP oscillations (i.e. synchronized activity) in the alpha and beta band (10–30 Hz), gamma band (30–100 Hz), and theta band (4–8 Hz; Bahramisharif et al., 2017; Bastos et al., 2018; van Ede et al., 2017; Honkanen et al., 2015; Howard et al., 2003; Lundqvist et al., 2016; Roux et al., 2012; Salazar et al., 2012).

The gamma band has been associated with sensory information held in working memory (Bastos et al., 2018; Honkanen et al., 2015; Howard et al., 2003; Roux et al., 2012) as well as spiking carrying sensory information (Lundqvist et al., 2016, 2018a). In fact, gamma power correlates with the number of objects held in working memory (Howard et al., 2003; Kornblith et al., 2016; Roux et al., 2012). The alpha and beta band has been associated with top-down information (e.g., task rules) and with inhibitory functions (discussed below). It is anti-correlated with gamma. The theta band may play a role in generating irregular bursts of gamma and spiking (see below). As we will see, the interaction between these rhythms and spiking has provided insight into top-down "executive" control that gates access to working memory.

To be clear, we are not suggesting that the classic model of persistent spiking is wrong. It is correct at a certain level of approximation, averaged spiking of single neurons. But a new look in more detail (e.g., on single trials) and on a network level has provided new insights. The results still point to a central role for spiking in working memory. It is just that the story is more complex than previously suspected. Is not that always the case?

Persistent Problems

We recently reviewed evidence for and against persistent spiking underlying working memory (Lundqvist et al., 2018b), so we will be brief here.

The evidence associating delay interval spiking with working memory maintenance is clear and unequivocal (e.g., Funahashi et al., 1989; Fuster, 1999; Fuster and Alexander, 1971; Goldman-Rakic, 1995; Miller et al., 1996; Pasternak and Greenlee, 2005; Romo et al., 1999). However, the evidence that spiking is persistent is less so. Virtually all prior studies averaged spiking over time, across trials, and often across neurons recorded in different sessions. Averaging masks the details of spiking activity (Stokes and Spaak, 2016). Single-trial analyses indicate spiking is typically sparse in real time (Kucewicz et al., 2017; Lundqvist et al., 2016, 2018a; Shafi et al., 2007; Stokes and Spaak, 2016).

Yes, there are examples in the literature of single neurons that seem to show persistent spiking on individual trial rasters. This suggests that at least some neurons show persistent activity. But the bulk of neurons spike sparsely in working memory delays, even when spiking is averaged across trials (Cromer et al., 2010; Fujisawa et al., 2008; Hussar and Pasternak, 2012; Shafi et al., 2007). A model that only explains the properties of a small percentage of the population is not complete. In addition, those examples are almost all from single-neuron studies in which investigators (necessarily) chose to only study neurons that seemed to show a property of interest (like delay activity spiking). That, plus single-neuron examples, are invariably "best of," means that they are hardly representative of the underlying population. Further, single-neuron studies typically optimize stimulus parameters for the neuron under study, thus optimizing neural activity. Under real-world conditions, however, only a tiny fraction of neurons may be operating under such ideal conditions. They are also not representative of the bulk of neurons contributing to a given function. Parsimony suggests that the whole population of neurons contribute to behavior, not just a select few operating under ideal conditions.

We are not saying that there is anything wrong with the approaches described above. They were and are essential for identifying constituent neural mechanisms (like delay activity). However, whether spiking activity is persistent versus sparse is a different level of question. It is one of how neural populations and the circuits they form contribute. This requires an approach in which neurons are sampled more randomly and in the context of the activity of other neurons so that network properties can be deduced. For this level of question, multiple-electrode studies that record activity of dozens to hundreds of neurons are better suited than single-neuron recording (Lundqvist et al., 2018b; Miller and Wilson, 2008).

But if single neurons do not show persistence, is it possible that it can be seen on the level of populations of neurons? This possibility rests on the assumption that single neurons spike asynchronously (i.e., at different times). When combined across different neurons, spiking "fills" time, producing persistence at the population level. To test this, one needs to measure activity in local networks, not just single neurons. This can involve analysis of multiple simultaneously recorded neurons as well as LFPs, which provide a measure of coordinated activity of neurons within a few hundred micrometers. We recently applied this approach to examine delay activity across seven frontal cortical areas (dorsolateral PFC, ventrolateral PFC, frontal eye fields [FEFs], dorsal premotor cortex, 8A, 8B, and the supplementary motor area or anterior cingulate cortex). As expanded below, this indicated that local populations of neurons are not asynchronous. Instead, there are sparse and coordinated bursts of spiking (Bastos et al., 2018; Lundqvist et al., 2016, 2018a).

Of course, one could posit that, if you combine enough neurons across a wide enough expanse of cortex, one can fill time with spikes. In other words, it could be that activity is persistent when combined across highly distributed networks. However, in order to evaluate extant models, the local network is critical. Much of the brain's computations take place on a local level. The cortex is thought to be organized into local, recurrently connected clusters with shared tuning properties (Constantinidis et al., 2001; Kritzer and Goldman-Rakic, 1995), and persistent activity is typically modelled using local recurrent connectivity (Amit and Brunel, 1997; Compte et al., 2000; Durstewitz et al., 2000).

Nonetheless, we can consider global cortical activity by using techniques like electroencephalogram (EEG) and fMRI. This has revealed that, for extended periods of time, information held in working memory cannot be decoded from global activity. However, when the cortex is "pinged" by a task-irrelevant stimulus or by transcranial magnetic stimulation, the network "rings" back with the information held in working memory (Rose et al., 2016; Sprague et al., 2016; Stokes et al., 2013; Wolff et al., 2017). This suggests that the working memory can be held in the absence of persistent spiking.

Finally, most of the evidence that working memory involves simply maintaining ensembles in a persistent active state comes from relatively simple tasks in which a single item must be retained over a "blank" delay interval (with no intervening distractions, further additions to working memory, etc.). That favors evidence for persistence spiking by "protecting" it from events that might disrupt it. When an interruption occurs, for example, by having the animal focus briefly on another task, delay activity can be disrupted for 100s of milliseconds without any loss of the working memory items (Spaak et al., 2017; Watanabe and Funahashi, 2014). It is possible, in principle, that working memory items could be switched in and out of long-term memory to bridge these gaps. But that would still require maintenance of an index to the information in long-term memory.

Another related issue is the stability of the neural code underlying working memory. New sensory inputs can change the neural population code carrying working memories. This can be evaluated by testing whether a decoder trained on activity at one time in the trial can decode information at other times. If not, then there has been a change in code. Even without intervening inputs, the population code changes over the memory delay (Meyers et al., 2008; Spaak et al., 2017; Stokes et al., 2013). This argues against a model of working memory in which an ensemble is activated by a sensory input and then kept in that active state. Instead, working memory representations are dynamic and change over time. It is possible, however, to find a linear combination of neurons that will maintain a stable code, "a stable subspace" (Murray et al., 2017). However, this has been demonstrated with blank delays without additional inputs or distractions. Decoders trained on time before additional inputs do not perform well following it (Parthasarathy et al., 2017).

Further, computational modeling of persistent activity using attractor dynamics suggests its limitations. Attractor dynamics are network dynamics dominated by neurons with persistent spiking. Different attractor states correspond to unique patterns of activity corresponding to different items in working memory. As long as the state is maintained, the memories are held. The problem is that attractor states are not stable when they are perturbed. They can be disrupted by a distracting input or by adding additional information to working memory. For this same reason, they have difficulty storing more than one working memory at a time. Bump attractor models, originally proposed for visuospatial working memory, can store multiple locations if there is no overlap in their neural representations, that is, if the working memories are held by essentially different networks (Almeida et al., 2015; Edin et al., 2009). But if there is overlap, the attractor states for different working memories tend to meld into one. This is problematic for the overlapping representations seen in the PFC (the cortical area most associated with working memory), at least for non-spatial information (Rigotti et al., 2013; Warden and Miller, 2010). Any universal model of working memory needs to deal with overlapping representations. Otherwise, it is only a special-case model.

What Is the Alternative?

An alternative is a hybrid attractor-dynamic and synaptic model. Rather than persistent spiking, there are brief, sparse, bursts of spiking. Working memories are held between spiking by spiking-induced changes in synaptic weights, "impressions" left in the network (Lundqvist et al., 2011, 2012; Mongillo et al., 2008; Sandberg et al., 2003; Stokes, 2015). Wang, Goldman-Rakic, and colleagues showed that spiking in the PFC can produce fast synaptic enhancement that lasts hundreds of milliseconds (Wang et al., 2006). In fact, the enhancement depends on sparse, bursty spiking. Not only is this metabolically less expensive, it also mitigates many of the problems of persistent attractor states. Synaptic weights are less prone to interference. Because the time spent in active attractor states is kept to a minimum, the working memories are less prone to disruption from, e.g., a new sensory input. Multiple items can be simultaneously held by multiplexing in time their brief bouts of activity. In other words, by having different ensembles active at different times, the attractor states do not interfere with each other (e.g., Siegel et al., 2009).

For example, in the synaptic attractor model (SAM), ensembles have inhibitory connections with other ensembles (Lundqvist et al., 2011), a feature shared by classic models of working memory (Amit and Brunel, 1997; Goldman-Rakic, 1996). Each attractor state has a limited lifetime. Thus, they are semi-stable and shut others down temporarily. The result is that each working memory item is expressed in brief bouts of spiking. Based on known biophysics, the SAM predicts that, in absence of bottom-up sensory inputs, networks oscillate in the alpha and beta band (10–30 Hz), only occasionally spiking. When a bottom-up sensory input activates an ensemble, it temporarily oscillates in a gamma state (>30 Hz) and gives off a short burst of elevated spiking before inhibition reverts it back to the alpha and beta state and reduced spiking. The gamma bursts may be linked to underlying theta rhythms (Canolty et al., 2006; Voytek et al., 2015; Watrous et al., 2015). This could organize time-multiplexing of items (Bahramisharif et al., 2017; Fuentemilla et al., 2010; Herman et al., 2013).

The spiking induces temporary (<1 s) changes in synaptic weights, perhaps via calcium dynamics (Lundqvist et al., 2011, 2016; Mongillo et al., 2008; Wang et al., 2006). Therefore, both spiking and short-term plasticity are thought to be mechanisms for working memory storage. Brief, irregular bursts of spiking and gamma during the memory delay are needed to occasionally refresh the synaptic weight changes so that the working memories can be maintained beyond the lifetime of the synaptic weight changes.

In the model, the refresh rate is responsible for the limited capacity of working memory (an average of four items; Awh et al., 2007; Buschman et al., 2011; Cowan, 2010; Luck and Vogel, 1997). If too many items are simultaneously held, the requirement to refresh the synapses causes a buildup of interference due to competition for the limited time available for the refresh (Lundqvist et al., 2011; Mi et al., 2017). For this reason, the gamma burst rate increases with working memory load (Lundqvist et al., 2016). Schizophrenic patients have lowered working memory capacity and do not demonstrate the load-dependent changes in gamma (Basar-Eroglu et al., 2007) observed in healthy subjects (Howard et al., 2003; Roux et al., 2012).

We tested this model by analyzing LFP and spiking from seven cortical areas (dorsolateral and ventrolateral PFC, the frontal eye fields, dorsal premotor cortex, areas 8A and 8B, and the

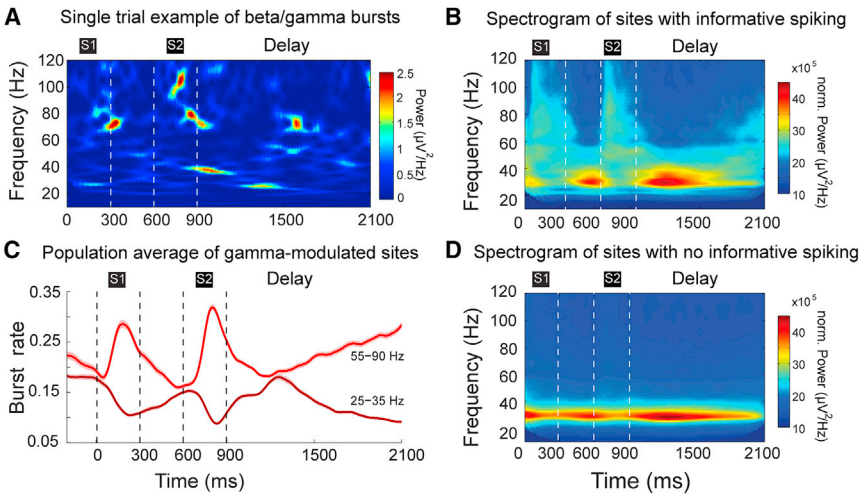

Figure 1. Gamma and Beta Bursts Underlie Working Memory

(A) A single-trial example of LFP power in time and frequency. Two stimuli were presented (S1 and S2) and later tested following a delay. Narrow bursts of power in the beta and gamma bands are evident both during cue processing and delay.

(B–D) LFP data from sites that contained spikes that carried information about the presented cue (B) versus those that did not (D) are shown. Only sites containing informative spiking (D; population average) showed modulation of beta and gamma. This effect remained after controlling for differences in spike rate between informative (B) versus non-informative (C) sites.

(C) On gamma-modulated sites, the beta and gamma burst rates are mirror images of each other. Gamma bursting increases during stimulus presentation and towards the end of the delay, and beta does the opposite.

(D) On sites without informative spiking, only beta is task modulated and less so.

Modified from Lundqvist et al. (2016).

supplementary motor area or anterior cingulate cortex) of monkeys performing several different working memory tasks (Bastos et al., 2018; Lundqvist et al., 2016, 2018a). These tasks involved both spatial and non-spatial working memory and different working memory loads (1–3). Across all these different tasks and areas, spiking that carried information about the sensory inputs to be held in working memory was highly associated with brief bursts of narrow-band gamma oscillations, especially during the encoding of sensory information into working memory (Figures 1A–1C; Lundqvist et al., 2016). During such gamma bursts, spiking was elevated and more informative about the contents of working memory than spiking outside the bursts (Lundqvist et al., 2018a). In fact, at recording sites where spiking did not carry working memory information, there was little or no gamma bursting (Figure 1D). Interleaved with the gamma bursts were brief bursts of beta and bursts that were not associated with spiking carrying working memory contents. During the memory delays, the gamma bursts occurred at a lower rate but still above the baseline rate (Bastos et al., 2018; Lundqvist et al., 2016, 2018a). This is consistent with the model prediction that gamma bursts and spikes are needed to refresh synaptic weight changes. The gamma bursting and associated spiking increased near the end of the delay, around the time working memories needed to be "read out" (Figure 1C).

Importantly, gamma bursts and alpha and beta bursts were anti-correlated, like mirror images of each other (Figure 1C). This was task related, only appearing at recording sites where spiking reflected the contents of working memory (Bahramisharif et al., 2017; Lundqvist et al., 2016, 2018a). The task-related anti-correlation between gamma bursts and alpha and beta bursts intrigued us. It occurred to us that it could be a mechanism for controlling working memory storage. Gamma is associated with the spiking that holds sensory inputs in working memory. If it has a push-pull relationship with alpha and beta, then gamma (and hence working memory storage) can be turned on and off by lowering and raising alpha and beta, respectively. For example, turning down alpha and beta would allow gamma to be expressed and sensory inputs to be encoded in working memory. Turning up alpha and beta would turn down gamma and thus clear out the working memory storage.

What about Alpha and Beta?

The above implies that alpha and beta have an inhibitory role in working memory. In visual cortex, inhibition has been linked with alpha (8–12 Hz; Haegens et al., 2011; Jensen and Mazaheri, 2010). In prefrontal and motor cortex, inhibition is more often linked with beta (15–30 Hz). However, several studies report power modulation that spans both the alpha and beta bands (Bastos et al., 2018; van Ede et al., 2011). Thus, we will group these bands together as they seem to have similar functions: providing inhibition. One exception is in parietal cortex, where lower beta has been associated with working memory maintenance (Kopell et al., 2011; Salazar et al., 2012).

Motor planning has similarities to working memory control and may have shared evolutionary origins (Chatham and Badre, 2015). In fact, motor beta and gamma have very similar behavioral correlates as working memory beta and gamma. Beta is elevated when a movement is being withheld (Donoghue et al., 1998; Feingold et al., 2015; Jha et al., 2015; Zhang et al., 2008). During movement, beta wanes and gamma waxes. Beta is then elevated after movement (Feingold et al., 2015) as if the motor plan was being cleared out. Similarly, there was increased beta in the PFC after the end of a trial, once working memories are no longer relevant (Lundqvist et al., 2018a). In fact, this effect was selective to recording sites where working memory information was held during the trial. Alpha/beta may also play a role in protecting working memory from distractors (Bonnefond and Jensen, 2012). Across virtually all of sensory cortex, gamma is associated with sensory processing and beta is anti-correlated with gamma (Bauer et al., 2006; David et al., 2015; van Ede et al., 2011; Fisch et al., 2009; Fontolan et al., 2014; Zhang et al., 2008).

Inhibition is central to executive control and so is the knowledge about what needs to be controlled (Miller and Cohen, 2001). Correspondingly, beta has also been associated with the top-down information. Task rules are reflected in different

patterns of beta synchrony in PFC (Buschman et al., 2012) and visual cortex (Richter et al., 2018) as if beta was helping form ensembles for the rules. Such content-specific "beta ensembles" have also been found for other types of top-down information, like learned categories (Antzoulatos and Miller, 2016; Stanley et al., 2018; Wutz et al., 2018). Thus, with the spatiotemporal pattern of beta changing with top-down information, beta's inhibitory effects can act selectively and direct the flow of sensory information.

Support for this comes from numerous studies showing that attention to sensory inputs results in increased gamma and increased alpha and beta occurs for modalities or locations that are unattended (Buffalo et al., 2011; Fries et al., 2001; Haegens et al., 2011; Jensen and Mazaheri, 2010; van Ede et al., 2017; Leszczynski et al., 2017; Popov et al., 2017; Wolff et al., 2017). A magnetoencephalography (MEG) study in humans also showed that the alpha and beta in sensory cortex were anti-correlated with the locus of attention (and with gamma) and were under top-down control from frontal cortex (Popov et al., 2017). The alpha and beta were also anti-correlated with behavioral reaction time, indicating its functional relevance.

Thus, we propose dual roles for beta: inhibition and formation of ensembles for top-down information. We hypothesize that the inhibitory role for beta is a mechanism acts locally, at the level of cortical columns (Bastos et al., 2018). This local inhibition is akin to the role proposed for alpha in sensory cortex (Jensen and Mazaheri, 2010). In addition, beta rhythms have been proposed to be ideally suited for flexibly generating neural ensembles (Kopell et al., 2011; Spitzer and Haegens, 2017), with the beta rhythmic networks reaching down to the level of individual cells (Dann et al., 2016). These large-scale neural ensembles, we propose, contain the top-down knowledge required to locally deliver inhibition, and thus executive control, where and when it is needed.

In correspondence with their roles in top-down versus bottom-up functions, beta and gamma have also been associated with feedback and feedforward cortical processing. In a study using large-scale electrocorticography, Bastos et al. (2015a) recorded from eight different visual areas simultaneously as monkeys performed a visual attention task. A corticocortical motif emerged by analyzing all pairs of areas in relation to their anatomical pattern of feedforward and feedback connectivity. Gamma oscillations were shown to flow up the visual cortical hierarchy in a bottom-up direction. Beta oscillations flowed down the hierarchy in the top-down direction. A similar functional hierarchy was then subsequently discovered in the human visual system with MEG recordings (Michalareas et al., 2016). Causal evidence also supports these findings. Electrical micro-stimulation in V1 causes increases in gamma power in V4, an area downstream from V1 and in receipt of feedforward connections. Micro-stimulation in V4 causes increases in alpha power in V1 (van Kerkoerle et al., 2014).

Note that bottom-up gamma is not inconsistent with the idea that top-down attention often enhances gamma power and inter-area synchrony (Bastos et al., 2015a; Bosman et al., 2012; Buschman and Miller, 2007; Fries et al., 2001; Gregoriou et al., 2009). Top-down attention is often conceptualized as a "spotlight" that turns up the gain on behaviorally relevant sensory representations (Desimone and Duncan, 1995). Thus, sensory enhancement of attended items also enhances gamma. At the same time, gamma enhancement can be controlled by beta rhythms (Lee et al., 2013). Richter and colleagues examined the trial-by-trial pattern of top-down Granger causality from parietal to visual cortex in beta with the bottom-up Granger causality from V1 to V4 in gamma (Richter et al., 2017). The strength of top-down (parietal to visual cortex) beta synchrony predicted the strength of bottom-up (V1–V4) gamma synchrony.

Plugging this into what we suggested above, the idea is that top-down information carried by alpha and beta rhythms could inhibit the expression of bottom-up information carried by gamma rhythms and perhaps even regulate the precise patterns of gamma synchrony that enable corticocortical communication (Fries, 2015). But how do these rhythms interact on a micro-circuit level? The answer seemed to lie in interactions between cortical layers.

Beta in Deep-Layer Cortex Interacts with Gamma in Superficial-Layer Cortex

The cerebral cortex has laminar organization. Layer 4 is the input layer (Felleman and Van Essen, 1991; Gilbert and Wiesel, 1983; Rockland and Pandya, 1979). Although the correspondence is not perfect (biology never is), the superficial layers (layers 2 and 3) largely contain the feedforward-projecting neurons that carry sensory information anteriorly and the deep layers (layers 5 and 6) contain the feedback-projecting neurons that carry the top-down information posteriorly in cortex (Markov et al., 2014). Gamma and beta rhythms are emphasized in different cortical layers. In visual cortex, gamma is more prominent in superficial and middle layers and alpha and beta are more prominent in deep layers (Bollimunta et al., 2008; Buffalo et al., 2011; Maier et al., 2010; Smith et al., 2013; Xing et al., 2012).

To determine whether this was also true in frontal cortical areas associated with working memory, we recorded with "laminar" electrodes in animals performing three different working memory tasks (Bastos et al., 2018). Laminar electrodes have multiple contacts along the shaft and thus allow recording from all cortical layers simultaneously.

Frontal cortex gamma power and cue-related information peaked in superficial layers and alpha and beta peaked in deep layers (Figure 2A). Working memory delay interval spiking was also stronger in superficial layers (Figure 2B). This corresponds with our observations about gamma and beta rhythms. The superficial layers are the feedforward layers that carry bottom-up sensory inputs up the cortical hierarchy. Thus, it is where we would expect to find more bottom-up gamma and spiking carrying sensory information held in working memory. In sensory areas, bottom-up gamma and informative spiking is typically only elevated for the duration of sensory stimulation (Buffalo et al., 2011; Fries et al., 2001). In prefrontal cortex, bursts of spiking and gamma also appear over the delay interval. This could be the result of longer time synaptic integration constants (Murray et al., 2014) brought about by superficial-layer lateral excitatory connections (Goldman-Rakic, 1996) and more synaptic spines on pyramidal cells (Elston, 2000). Likewise, it makes sense that beta would be stronger in deep layers. Beta is associated with top-down information. The deep layers are the

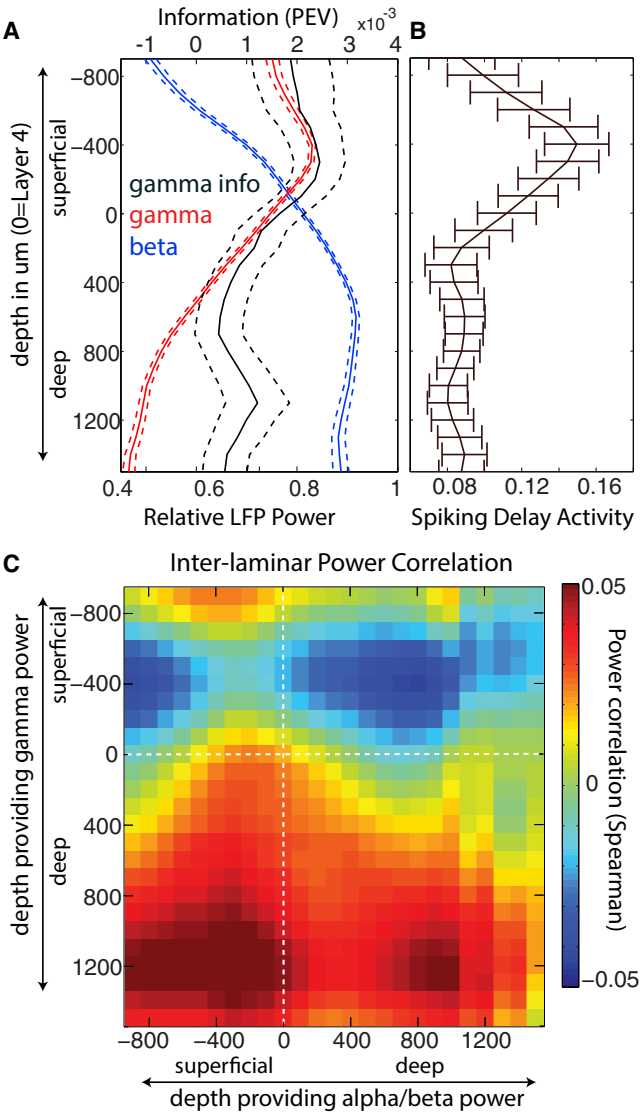

Figure 2. Laminar Organization of Gamma/Beta Rhythms and Delay Activity

(A) Gamma power and alpha and beta power are segregated into distinct layers. Gamma power peaks 400 μm above layer 4, whereas alpha and beta power peaks at 600 μm below layer 4. Gamma bursts in superficial, but not deep, layers carry significant information about the cued item during the working memory delay period (quantified by the percent explained variance [PEV] statistic). Beta bursts do not carry significant information during the delay (not shown). Dotted lines are ±1 SEM across sessions (N = 60).

(B) Spiking activity, quantified by multiunit change from baseline (a.u.) during the delay period, is strongest in superficial layers. The pattern of laminar pattern of delay activity correlates strongly with gamma and is strongly anti-correlated with alpha and beta.

(C) Correlation map between gamma and beta power across layers. Deep-layer beta power is anti-correlated with superficial layer gamma power during the working memory delay.

From Bastos et al. (2018).

feedback layers that can carry top-down information from frontal cortex down the cortical hierarchy.

The pattern of influence between beta and gamma suggested a laminar-rhythmic infrastructure for control of working memory storage. Granger causality is a statistical measure of time series prediction that is indicative of functional connectivity (Bressler and Seth, 2011). It indicated that deep-layer beta oscillations regulated superficial-layer beta. The phase of deep-layer beta oscillations, in turn, modulated the amplitude of superficial gamma (Bastos et al., 2018; Canolty et al., 2006; Colgin et al., 2009; Lakatos et al., 2005; Spaak et al., 2012). Importantly, the power of deep-layer beta was inversely correlated with superficial-layer gamma, consistent with an inhibitory role for beta (Figure 2C). Thus, coupling between deep and superficial layers may serve a control function. Increasing deep-layer alpha and beta would increase superficial layer beta. Superficial layer beta would, in turn, suppress gamma and thus the expression of bottom-up sensory information in superficial layers. This would prevent the encoding of sensory information in working memory. Conversely, if deep-layer beta is reduced, there would be decreased coupling to superficial-layer beta. That would release gamma from inhibition, allowing its expression and the encoding of bottom-up information into working memory. Indeed, we found that the strength of deep-layer beta coupling to superficial-layer gamma was reduced during the working memory delays compared to the pre-cue baseline period (Bastos et al., 2018).

Mechanisms of Gamma/Beta Interplay

To understand how the interplay between gamma and beta gives rise to working memory control, it is important to consider their neurophysiological origins. Here, we provide a short summary (for detailed reviews, see Buzsáki and Wang, 2012; Fries, 2015; Wang, 2010).

Excitatory (E) and inhibitory (I) cells are densely interconnected in cortex. Fast (greater than 10 Hz) rhythms can be generated in cortex through recurrent inhibition between E cells and a variety of classes of I cells. Fast-spiking (FS) I cells are a key player. They provide the feedback inhibition necessary to shut down activity and create an oscillation. Once the inhibition wears off it creates a window for the E cells to fire. The inhibitory time constants determine the spacing of these time windows and thus the rhythmic frequency. Other relevant factors that determine the length of the oscillatory cycle are the input strength to the network, the pattern of connectivity between the E and I cells, and the leak currents (Brunel and Wang, 2003). This mechanism has been termed "PING," pyramidal interneuron network gamma, because it was originally conceived as a mechanism for gamma (Amit and Brunel, 1997; Brunel and Wang, 2003; Whittington et al., 2000). However, it can also generate beta-rhythmic ensembles (Lee et al., 2013; Lundqvist et al., 2011). We should note that our hypotheses about the role of beta in working memory do not depend on how beta is generated. We offer this as one possible mechanism; there are other models for beta generation (Sherman et al., 2016).

There could be two separate PING mechanisms in the superficial and in deep layers: stronger gamma in superficial and stronger beta in deep layers resulting from different classes of I cells in superficial versus deep layers with different time constants and/or the greater number of FS cells in superficial layers. The observed push-pull interaction between superficial gamma and deep beta could be generated by reciprocal

inhibitory connections between the two PING networks (Lee et al., 2013).

The PING mechanism relies on strong excitation. For the gamma band, this drive is the sensory stimulation itself. In visual cortex, this generates strong, oscillatory gamma in response to sensory input, which ceases when the stimulus is removed (Bastos et al., 2015a; Bosman et al., 2012; Brunet et al., 2015; Fries et al., 2008). In the PFC, gamma is more bursty and variable, e.g., the center frequency varies and the bursts are sparse (Lundqvist et al., 2016). This may be because the PFC integrates inputs from many cortical and subcortical areas. Thus, external sensory drive will have less of an impact on its overall excitation. Also, likely due to an enhanced number of excitatory connections on PFC cells (Elston, 2000), more lateral excitatory connectivity (Goldman-Rakic, 1996), longer intrinsic time constants (Murray et al., 2014), and synaptic mechanisms (Wang et al., 2006), PFC networks are able to produce (bursty) gamma even in the absence of sensory stimuli.

The relationship between sensory input and rhythms is opposite for beta. Beta is more prominent in the absence of sensory drive (and, in somatomotor cortex, the absence of motor movement). Deep-layer beta may be generated by a PING mechanism with excitatory drive provided via thalamocortical (Ketz et al., 2015) and/or basal ganglia (Chatham and Badre, 2015) loops that are self-sustaining in the absence of external inputs. Thus, beta is strong in the absence of sensory inputs, during planning, task set preparation, etc. Competition between the beta and gamma assemblies could control the "tuning" of the network to either internal (in beta) or external (in gamma) information (Brincat and Miller, 2016; Buschman and Miller, 2007).

Interplay between Gamma and Beta during Working Memory Control

To test whether this interplay between beta and gamma correlates with the control of working memory, we used a sequence-matching task (Lundqvist et al., 2018a). Animals held sequences of two objects in working memory and then had to judge whether a subsequent test sequence was a match. The advantage to this task is that it has multiple decision points. Animals have to determine whether each object is a match both in identity and order. This affords more opportunity to examine working memory control than a typically working memory task, which only involves remembering one stimulus and making one decision that co-occurs with a motor action.

This analysis revealed that shifts in the balance between beta and gamma and spiking did, in fact, correlate with working memory control (Lundqvist et al., 2018a). In anticipation of having to use a given object for the match decision (e.g., the first sample object for judging the first test object or the second for the second), there was reduced beta bursting along with an increase in gamma bursting and spiking information about the specific anticipated object. When an object held in working memory was no longer needed, beta increased and gamma decreased together with spiking conveying information about that object. Further, deviations from "correct" beta and gamma dynamics predicted not only a forthcoming error but what kind of error the animal would make. For example, if the animal was going to mistakenly respond "match" to a non-matching sequence, the temporal dynamics of and balance between gamma and beta bursting looked like that on a match trial instead of non-match trial. We could also tell whether the animals made the wrong decision to the first or second test object. In short, shifts in the balance between beta and gamma correlated with working memory control processes; errors in the balance predicted upcoming behavioral errors.

Cortical Gradients

So far, our discussion has emphasized the PFC. Delay activity spiking, however, is a widespread cortical phenomenon (Fuster, 2015). But how widespread has recently generated vigorous debate (Christophel et al., 2017; Leavitt et al., 2017). For example, Dotson and Gray recorded spiking activity from 42 cortical areas (Dotson et al., 2018). Delay activity was widespread but also showed gradients. In V1, delay activity was mostly decreased spiking in the delay relative to baseline, suggesting synaptic adaptation. This could also be a consequence of top-down signaling from higher cortical areas (van Kerkoerle et al., 2017). Other studies in early sensory areas also showed weak or non-existent delay activity spiking compared to higher-order cortical areas (Haller et al., 2018; Leavitt et al., 2017; Mendoza-Halliday et al., 2014). Interestingly, at the other extreme of cortical processing, in motor cortex, there is also little delay activity (Dotson et al., 2018; Haller et al., 2018). In between, there is higher-order association cortex (including PFC, posterior parietal cortex, and temporal cortex) rich in delay activity (Dotson et al., 2018; Fuster, 1990; Haller et al., 2018; Leavitt et al., 2017; Sigala, 2009; Woloszyn and Sheinberg, 2009). These areas are highly interconnected (Markov et al., 2013). They are also the cortical areas where top-down and bottom-up information reaches apex (Brincat et al., 2018; Siegel et al., 2015) and thus could support domain-general cognitive operations (Haller et al., 2018).

These higher order, delay-activity-rich areas share several aspects of laminar circuitry. They have a balance in the soma size and corticocortical output connectivity between superficial versus deep layers (Goulas et al., 2018). In contrast, motor output structures have a large laminar asymmetry in soma size (larger deep layer neurons) and predominant layer of cortical output (deep layers). Low-level sensory cortex features a highly differentiated and dense laminar circuit, emphasizing superficial layer soma size, and most corticocortical outputs originate from superficial layers.

We hypothesize that low-level sensory and motor cortex is not ideal for working memory representation and control as a result of their local circuitry. Sensory areas have a relative emphasis on the superficial layers (Zaldivar et al., 2018), where inputs can be richly encoded with gamma but lack the control element from deep layers. Motor areas emphasize deep layers (along with a predominance of beta), where outputs to motor structures can be gated but have a relatively poor superficial layer circuitry (Goulas et al., 2018). Association cortices lie in between. They have a relative balance between superficial and deep circuitry (Goulas et al., 2018) better suited for both representation and control of activity. In addition, there are other neuroanatomical gradients that also change from early to higher-order cortex, such as spine density and lateral connectivity (Elston, 2000;

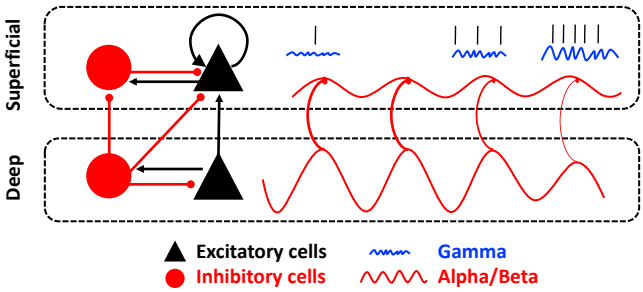

Figure 3. A Model of Working Memory
Denoted by two rectangular, dashed boxes, two cortical compartments, superficial and deep, are made up of densely interconnected excitatory pyramidal (black) and inhibitory (red) interneurons. Inhibitory connections are line segments with a red, rounded end, and excitatory connections are line segments with a black, arrow end. Two separate PING networks in superficial versus deep layers are responsible for generating gamma in superficial layers and beta in deep layers (sustained by connections to thalamus and basal ganglia; not shown). The looping arrow returning on itself in the superficial layers represents the recurrent connectivity found within layer 3 pyramidal cell networks in prefrontal cortex. The sinusoidal red line in deep layers reflects beta oscillations and their driving influence on superficial beta oscillations. Beta oscillations are phase amplitude coupled with gamma oscillations (blue squiggly lines), and these gamma oscillations organize delay-period spiking, representing working memory content (straight black marks). Spiking activity inside gamma bursts is more informative than outside. Over time, moving from left to right in the figure, the deep beta reduces in power and releases inhibition onto the superficial layers. This results in enhanced superficial gamma and spiking, i.e., enhanced maintenance of working memory, as is seen when transitioning between baseline to working memory task performance. The reversed process (enhancement of deep layer beta and enhanced suppression of superficial layer gamma and spiking) would "clear out" the contents of working memory, as seen at the end of the trial, or when working memory contents are no longer needed.
From Bastos et al. (2018).

Goldman-Rakic, 1996). Both increase up the hierarchy, making cells more intrinsically excitable and integrative (Murray et al., 2014; Wasmuht et al., 2018). The relative balance between specific inhibitory cell populations also changes (Kim et al., 2017) and could impact circuits for working memory (Wang and Yang, 2018). It will be interesting to explore, in further work, which exact circuit elements enable higher-order cortex to sustain working memory.

Putting It All Together: A Model for Volitional Control of Working Memory

Figure 3 illustrates our model. It shares many aspects with previously proposed circuits for visual sensory function (Bastos et al., 2015b; Mejias et al., 2016). Spikes encode and help maintain information in working memory. Top-down information is associated with beta in deep (feedback) cortical layers (red wave). Bottom-up information is associated with gamma in superficial (feedforward) layers (blue waves). The central idea is that (top-down) deep-layer beta regulates the expression of (bottom-up) gamma in superficial layers, thus gating the access of sensory information to working memory and controlling its maintenance. Alpha/beta and gamma oscillations can be below the threshold for spiking, but they drive membrane potentials toward and sometimes over spike thresholds, which is why there tends to be more spiking on the depolarizing phases of oscillations (Siegel et al., 2009).

Both superficial and deep layers of cortex are comprised of networks of deeply interconnected excitatory pyramidal (black) neurons and inhibitory (red) interneurons. Deep-layer beta is unidirectionally coupled to superficial layer beta. In turn, superficial-layer beta suppresses superficial-layer gamma oscillations. Note that the middle and deep layers of PFC are reciprocally connected with the mediodorsal nucleus of the thalamus, with layer 4 receiving thalamic input and layers 5 and 6 sending output to the thalamus (Giguere and Goldman-Rakic, 1988). Working memory delay interval spiking is prominent in the medial dorsal (MD) thalamus (Watanabe and Funahashi, 2004). Beta-band coherence has been reported between PFC and thalamus during working memory maintenance (Parnaudeau et al., 2013). Optogenetics suppression of MD thalamus suppresses cortical delay activity (Schmitt et al., 2017). Thus, the modulatory role of beta in the deep layers for working memory control might be in part regulated by the thalamocortical loop.

To encode information in working memory, deep-layer beta power and/or its coupling to superficial-layer beta weakens. This disinhibits the recurrent excitation of layer 2 and layer 3 neurons (as indicated by the loop arrow) generating bursts of gamma. The gamma allows expression of spiking carrying bottom-up sensory inputs. The balance between beta and gamma can regulate the level of gamma bursting in the memory delay needed to occasionally refresh the synaptic weight changes that help maintain the working memories. During working memory readout, beta is once again relaxed, allowing the increased gamma bursting and the ramp-up of spiking often seen near the end of memory delays (Hussar and Pasternak, 2010; Roesch and Olson, 2005). Increased spiking is needed so that working meories can acquire control of behavior. Keeping gamma bursting and spiking at a lower level earlier in the delay interval may prevent working memories from prematurely acquiring that control. To clear out working memory, beta power and coupling increases. This suppresses gamma and the spiking that was maintaining the working memory.

Summary and (Many) Open Questions

Recent studies continue to indicate that memory delay spiking plays a critical mechanism for maintaining information in working memory. But they also indicate that there is more going on than a simple persistence of spiking. Instead, there are brief bursts of spiking and associated gamma bursting that reflect activation and reactivation of the attractor states of the neural ensembles for the working memory memoranda. The spiking could cause temporary changes in synaptic weights that carry the working memories between spiking. This combination of spiking and synaptic weight changes solves many of the problems with persistent spiking. It is metabolically less expensive and makes the memories more robust to interference. It allows multiple items to be held in working memory by "juggling" their activations in time. This new perspective is part of mounting evidence that the neural basis of cognition is not continuous (Buschman and Miller, 2010; Fiebelkorn et al., 2018; Helfrich et al., 2018; Landau and Fries, 2012; Schroeder et al., 2010; VanRullen, 2016).

Sparse spiking also leaves room for rhythmic interplay between oscillations of different bands: gamma, alpha, and beta.

Beta is associated with top-down information and seems to have an inhibitory role. Increasing beta decreases gamma and spiking and vice versa. Thus, the push-pull relationship (when beta is up, gamma is down and vice versa) may be the infrastructure for top-down, executive control of working memory storage. In short, beta can turn on and off the "faucet" of gamma-related working memory reactivations.

Our discussion has been focused on working memory and the higher cortical areas associated with working memory. But we have noted that there is a similar laminar distribution of gamma and beta as well as a similar push-pull relationship between them all over sensory and motor cortex. Thus, rather than playing a role in working memory only, this laminar interplay may be a cortical motif, a general mechanism by which the cortex, writ large, can control the inflow and processing of bottom-up sensory inputs via top-down knowledge (Bastos et al., 2012). This indeed fits with reports of widely distributed delay activity (Dotson et al., 2018; Fuster, 2015) and the idea that PFC could have more of a control function (Lara and Wallis, 2015; Miller and Cohen, 2001).

To be sure, this is just a beginning. Thus far, we have focused on the relatively simple processes of gating access to, and clearing out of, working memory. But working memory control is more than encoding and clearing information. It also involves manipulation. Information in working memory can be transformed, reordered, and sequenced, etc. This requires control at the level of individual ensembles, not just a general gating mechanism. Long-term memories can be loaded into working memory; we do not yet know whether the same rhythmic interplay underlies this. What we are proposing is the infrastructure by which volition acts. There is also, of course, the big question of the genesis of volition itself. Our model, as with any other (including the classic model of working memory storage), is just a starting point for more hypothesizing, further testing, and, if it is merited, updating.

ACKNOWLEDGMENTS

This work was supported by NIMH R37MH087027, ONR MURI N00014-16-1-2832, The Picower Fellows Program, and the MIT Picower Institute Innovation Fund. We thank S.L. Brincat, M. Halassa, P. Herman, D. Pinotsis, and A. Wutz for helpful discussions.

REFERENCES

Almeida, R., Barbosa, J., and Compte, A. (2015). Neural circuit basis of visuospatial working memory precision: a computational and behavioral study. J. Neurophysiol. *114*, 1806–1818.

Amit, D.J., and Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cereb. Cortex *7*, 237–252.

Antzoulatos, E.G., and Miller, E.K. (2016). Synchronous beta rhythms of frontoparietal networks support only behaviorally relevant representations. eLife *5*, e17822.

Awh, E., Barton, B., and Vogel, E.K. (2007). Visual working memory represents a fixed number of items regardless of complexity. Psychol. Sci. *18*, 622–628.

Bahramisharif, A., Jensen, O., Jacobs, J., and Lisman, J. (2017). Serial representation of items during working memory maintenance at letter-selective cortical sites. bioRxiv. https://doi.org/10.1101/171660.

Basar-Eroglu, C., Brand, A., Hildebrandt, H., Karolina Kedzior, K., Mathes, B., and Schmiedt, C. (2007). Working memory related gamma oscillations in schizophrenia patients. Int. J. Psychophysiol. *64*, 39–45.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., and Friston, K.J. (2012). Canonical microcircuits for predictive coding. Neuron *76*, 695–711.

Bastos, A.M., Vezoli, J., Bosman, C.A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J.R., De Weerd, P., Kennedy, H., and Fries, P. (2015a). Visual areas exert feedforward and feedback influences through distinct frequency channels. Neuron *85*, 390–401.

Bastos, A.M., Litvak, V., Moran, R., Bosman, C.A., Fries, P., and Friston, K.J. (2015b). A DCM study of spectral asymmetries in feedforward and feedback connections between visual areas V1 and V4 in the monkey. Neuroimage *108*, 460–475.

Bastos, A.M., Loonis, R., Kornblith, S., Lundqvist, M., and Miller, E.K. (2018). Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory. Proc. Natl. Acad. Sci. USA *115*, 1117–1122.

Bauer, M., Oostenveld, R., Peeters, M., and Fries, P. (2006). Tactile spatial attention enhances gamma-band activity in somatosensory cortex and reduces low-frequency activity in parieto-occipital areas. J. Neurosci. *26*, 490–501.

Bollimunta, A., Chen, Y., Schroeder, C.E., and Ding, M. (2008). Neuronal mechanisms of cortical alpha oscillations in awake-behaving macaques. J. Neurosci. *28*, 9976–9988.

Bonnefond, M., and Jensen, O. (2012). Alpha oscillations serve to protect working memory maintenance against anticipated distracters. Curr. Biol. *22*, 1969–1974.

Bosman, C.A., Schoffelen, J.-M., Brunet, N., Oostenveld, R., Bastos, A.M., Womelsdorf, T., Rubehn, B., Stieglitz, T., De Weerd, P., and Fries, P. (2012). Attentional stimulus selection through selective synchronization between monkey visual areas. Neuron *75*, 875–888.

Bressler, S.L., and Seth, A.K. (2011). Wiener-Granger causality: a well established methodology. Neuroimage *58*, 323–329.

Brincat, S.L., and Miller, E.K. (2016). Prefrontal cortex networks shift from external to internal modes during learning. J. Neurosci. *36*, 9739–9754.

Brincat, S.L., Siegel, M., von Nicolai, C., and Miller, E.K. (2018). Gradual progression from sensory to task-related processing in cerebral cortex. Proc. Natl. Acad. Sci. USA *115*, E7202–E7211.

Brunel, N., and Wang, X.-J. (2003). What determines the frequency of fast network oscillations with irregular neural discharges? I. Synaptic dynamics and excitation-inhibition balance. J. Neurophysiol. *90*, 415–430.

Brunet, N., Bosman, C.A., Roberts, M., Oostenveld, R., Womelsdorf, T., De Weerd, P., and Fries, P. (2015). Visual cortical gamma-band activity during free viewing of natural images. Cereb. Cortex *25*, 918–926.

Buffalo, E.A., Fries, P., Landman, R., Buschman, T.J., and Desimone, R. (2011). Laminar differences in gamma and alpha coherence in the ventral stream. Proc. Natl. Acad. Sci. USA *108*, 11262–11267.

Buschman, T.J., and Miller, E.K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science *315*, 1860–1862.

Buschman, T.J., and Miller, E.K. (2010). Shifting the spotlight of attention: evidence for discrete computations in cognition. Front. Hum. Neurosci. *4*, 194.

Buschman, T.J., Siegel, M., Roy, J.E., and Miller, E.K. (2011). Neural substrates of cognitive capacity limitations. Proc. Natl. Acad. Sci. USA *108*, 11252–11255.

Buschman, T.J., Denovellis, E.L., Diogo, C., Bullock, D., and Miller, E.K. (2012). Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. Neuron *76*, 838–846.

Buzsáki, G., and Wang, X.-J. (2012). Mechanisms of gamma oscillations. Annu. Rev. Neurosci. *35*, 203–225.

Canolty, R.T., Edwards, E., Dalal, S.S., Soltani, M., Nagarajan, S.S., Kirsch, H.E., Berger, M.S., Barbaro, N.M., and Knight, R.T. (2006). High gamma power

is phase-locked to theta oscillations in human neocortex. Science *313*, 1626–1628.

Chatham, C.H., and Badre, D. (2015). Multiple gates on working memory. Curr. Opin. Behav. Sci. *1*, 23–31.

Christophel, T.B., Klink, P.C., Spitzer, B., Roelfsema, P.R., and Haynes, J.-D. (2017). The distributed nature of working memory. Trends Cogn. Sci. *21*, 111–124.

Colgin, L.L., Denninger, T., Fyhn, M., Hafting, T., Bonnevie, T., Jensen, O., Moser, M.-B., and Moser, E.I. (2009). Frequency of gamma oscillations routes flow of information in the hippocampus. Nature *462*, 353–357.

Compte, A., Brunel, N., Goldman-Rakic, P.S., and Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb. Cortex *10*, 910–923.

Constantinidis, C., Franowicz, M.N., and Goldman-Rakic, P.S. (2001). Coding specificity in cortical microcircuits: a multiple-electrode analysis of primate prefrontal cortex. J. Neurosci. *21*, 3646–3655.

Cowan, N. (2010). The magical mystery four: how is working memory capacity limited, and why? Curr. Dir. Psychol. Sci. *19*, 51–57.

Cromer, J.A., Roy, J.E., and Miller, E.K. (2010). Representation of multiple, independent categories in the primate prefrontal cortex. Neuron *66*, 796–807.

Dann, B., Michaels, J.A., Schaffelhofer, S., and Scherberger, H. (2016). Uniting functional network topology and oscillations in the fronto-parietal single unit network of behaving primates. eLife *5*, e15719.

David, F., Courtiol, E., Buonviso, N., and Fourcaud-Trocmé, N. (2015). Competing mechanisms of gamma and beta oscillations in the olfactory bulb based on multimodal inhibition of mitral cells over a respiratory cycle. eNeuro *2*, ENEURO.0018-15.2015.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. Annu. Rev. Neurosci. *18*, 193–222.

Donoghue, J.P., Sanes, J.N., Hatsopoulos, N.G., and Gaál, G. (1998). Neural discharge and local field potential oscillations in primate motor cortex during voluntary movements. J. Neurophysiol. *79*, 159–173.

Dotson, N.M., Hoffman, S.J., Goodell, B., and Gray, C.M. (2018). Feature-based visual short-term memory is widely distributed and hierarchically organized. Neuron *99*, 215–226.e4.

Durstewitz, D., Seamans, J.K., and Sejnowski, T.J. (2000). Neurocomputational models of working memory. Nat. Neurosci. *3* (*Suppl*), 1184–1191.

van Ede, F., de Lange, F., Jensen, O., and Maris, E. (2011). Orienting attention to an upcoming tactile event involves a spatially and temporally specific modulation of sensorimotor alpha- and beta-band oscillations. J. Neurosci. *31*, 2016–2024.

van Ede, F., Jensen, O., and Maris, E. (2017). Supramodal theta, gamma, and sustained fields predict modality-specific modulations of alpha and beta oscillations during visual and tactile working memory. J. Cogn. Neurosci. *29*, 1455–1472.

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegnér, J., and Compte, A. (2009). Mechanism for top-down control of working memory capacity. Proc. Natl. Acad. Sci. USA *106*, 6802–6807.

Elston, G.N. (2000). Pyramidal cells of the frontal lobe: all the more spinous to think with. J. Neurosci. *20*, RC95.

Engle, R.W., Tuholski, S.W., Laughlin, J.E., and Conway, A.R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. J. Exp. Psychol. Gen. *128*, 309–331.

Feingold, J., Gibson, D.J., DePasquale, B., and Graybiel, A.M. (2015). Bursts of beta oscillation differentiate postperformance activity in the striatum and motor cortex of monkeys performing movement tasks. Proc. Natl. Acad. Sci. USA *112*, 13687–13692.

Felleman, D.J., and Van Essen, D.C. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex *1*, 1–47.

Fiebelkorn, I.C., Pinsk, M.A., and Kastner, S. (2018). A dynamic interplay within the frontoparietal network underlies rhythmic spatial attention. Neuron *99*, 842–853.e8.

Fisch, L., Privman, E., Ramot, M., Harel, M., Nir, Y., Kipervasser, S., Andelman, F., Neufeld, M.Y., Kramer, U., Fried, I., and Malach, R. (2009). Neural "ignition": enhanced activation linked to perceptual awareness in human ventral stream visual cortex. Neuron *64*, 562–574.

Fontolan, L., Morillon, B., Liegeois-Chauvel, C., and Giraud, A.-L. (2014). The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. Nat. Commun. *5*, 4694.

Fries, P. (2015). Rhythms for cognition: communication through coherence. Neuron *88*, 220–235.

Fries, P., Reynolds, J.H., Rorie, A.E., and Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. Science *291*, 1560–1563.

Fries, P., Womelsdorf, T., Oostenveld, R., and Desimone, R. (2008). The effects of visual stimulation and selective visual attention on rhythmic neuronal synchronization in macaque area V4. J. Neurosci. *28*, 4823–4835.

Fuentemilla, L., Penny, W.D., Cashdollar, N., Bunzeck, N., and Düzel, E. (2010). Theta-coupled periodic replay in working memory. Curr. Biol. *20*, 606–612.

Fujisawa, S., Amarasingham, A., Harrison, M.T., and Buzsáki, G. (2008). Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. Nat. Neurosci. *11*, 823–833.

Funahashi, S., Bruce, C.J., and Goldman-Rakic, P.S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J. Neurophysiol. *61*, 331–349.

Fusi, S., Miller, E.K., and Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. Curr. Opin. Neurobiol. *37*, 66–74.

Fuster, J. (2015). The Prefrontal Cortex (Academic Press).

Fuster, J.M. (1990). Inferotemporal units in selective visual attention and short-term memory. J. Neurophysiol. *64*, 681–697.

Fuster, J.M. (1999). Memory in the Cerebral Cortex: An Empirical Approach to Neural Networks in the Human and Nonhuman Primate (MIT Press).

Fuster, J.M., and Alexander, G.E. (1971). Neuron activity related to short-term memory. Science *173*, 652–654.

Giguere, M., and Goldman-Rakic, P.S. (1988). Mediodorsal nucleus: areal, laminar, and tangential distribution of afferents and efferents in the frontal lobe of rhesus monkeys. J. Comp. Neurol. *277*, 195–213.

Gilbert, C.D., and Wiesel, T.N. (1983). Functional organization of the visual cortex. Prog. Brain Res. *58*, 209–218.

Goldman-Rakic, P.S. (1995). Cellular basis of working memory. Neuron *14*, 477–485.

Goldman-Rakic, P.S. (1996). Regional and cellular fractionation of working memory. Proc. Natl. Acad. Sci. USA *93*, 13473–13480.

Goulas, A., Zilles, K., and Hilgetag, C.C. (2018). Cortical gradients and laminar projections in mammals. Trends Neurosci. Published online July 3, 2018. https://doi.org/10.1016/j.tins.2018.06.003.

Gregoriou, G.G., Gotts, S.J., Zhou, H., and Desimone, R. (2009). High-frequency, long-range coupling between prefrontal and visual cortex during attention. Science *324*, 1207–1210.

Haegens, S., Nácher, V., Luna, R., Romo, R., and Jensen, O. (2011). α-oscillations in the monkey sensorimotor network influence discrimination performance by rhythmical inhibition of neuronal spiking. Proc. Natl. Acad. Sci. USA *108*, 19377–19382.

Haller, M., Case, J., Crone, N.E., Chang, E.F., King-Stephens, D., Laxer, K.D., Weber, P.B., Parvizi, J., Knight, R.T., and Shestyuk, A.Y. (2018). Persistent neuronal activity in human prefrontal cortex links perception and action. Nat. Hum. Behav. *2*, 80–91.

Helfrich, R.F., Fiebelkorn, I.C., Szczepanski, S.M., Lin, J.J., Parvizi, J., Knight, R.T., and Kastner, S. (2018). Neural mechanisms of sustained attention are rhythmic. Neuron *99*, 854–865.e5.

Herman, P.A., Lundqvist, M., and Lansner, A. (2013). Nested theta to gamma oscillations and precise spatiotemporal firing during memory retrieval in a simulated attractor network. Brain Res. *1536*, 68–87.

Honkanen, R., Rouhinen, S., Wang, S.H., Palva, J.M., and Palva, S. (2015). Gamma oscillations underlie the maintenance of feature-specific information and the contents of visual working memory. Cereb. Cortex *25*, 3788–3801.

Howard, M.W., Rizzuto, D.S., Caplan, J.B., Madsen, J.R., Lisman, J., Aschenbrenner-Scheibe, R., Schulze-Bonhage, A., and Kahana, M.J. (2003). Gamma oscillations correlate with working memory load in humans. Cereb. Cortex *13*, 1369–1374.

Hussar, C., and Pasternak, T. (2010). Trial-to-trial variability of the prefrontal neurons reveals the nature of their engagement in a motion discrimination task. Proc. Natl. Acad. Sci. USA *107*, 21842–21847.

Hussar, C.R., and Pasternak, T. (2012). Memory-guided sensory comparisons in the prefrontal cortex: contribution of putative pyramidal cells and interneurons. J. Neurosci. *32*, 2747–2761.

Jensen, O., and Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. Front. Hum. Neurosci. *4*, 186.

Jha, A., Nachev, P., Barnes, G., Husain, M., Brown, P., and Litvak, V. (2015). The frontal control of stopping. Cereb. Cortex *25*, 4392–4406.

Just, M.A., and Carpenter, P.A. (1992). A capacity theory of comprehension: individual differences in working memory. Psychol. Rev. *99*, 122–149.

van Kerkoerle, T., Self, M.W., Dagnino, B., Gariel-Mathis, M.-A., Poort, J., van der Togt, C., and Roelfsema, P.R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. Proc. Natl. Acad. Sci. USA *111*, 14332–14341.

van Kerkoerle, T., Self, M.W., and Roelfsema, P.R. (2017). Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. Nat. Commun. *8*, 13804.

Ketz, N.A., Jensen, O., and O'Reilly, R.C. (2015). Thalamic pathways underlying prefrontal cortex-medial temporal lobe oscillatory interactions. Trends Neurosci. *38*, 3–12.

Kim, Y., Yang, G.R., Pradhan, K., Venkataraju, K.U., Bota, M., García Del Molino, L.C., Fitzgerald, G., Ram, K., He, M., Levine, J.M., et al. (2017). Brain-wide maps reveal stereotyped cell-type-based cortical architecture and subcortical sexual dimorphism. Cell *171*, 456–469.e22.

Kopell, N., Whittington, M.A., and Kramer, M.A. (2011). Neuronal assembly dynamics in the beta1 frequency range permits short-term memory. Proc. Natl. Acad. Sci. USA *108*, 3779–3784.

Kornblith, S., Buschman, T.J., and Miller, E.K. (2016). Stimulus load and oscillatory activity in higher cortex. Cereb. Cortex *26*, 3772–3784.

Kritzer, M.F., and Goldman-Rakic, P.S. (1995). Intrinsic circuit organization of the major layers and sublayers of the dorsolateral prefrontal cortex in the rhesus monkey. J. Comp. Neurol. *359*, 131–143.

Kucewicz, M.T., Berry, B.M., Kremen, V., Brinkmann, B.H., Sperling, M.R., Jobst, B.C., Gross, R.E., Lega, B., Sheth, S.A., Stein, J.M., et al. (2017). Dissecting gamma frequency activity during human memory processing. Brain *140*, 1337–1350.

Lakatos, P., Shah, A.S., Knuth, K.H., Ulbert, I., Karmos, G., and Schroeder, C.E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. J. Neurophysiol. *94*, 1904–1911.

Landau, A.N., and Fries, P. (2012). Attention samples stimuli rhythmically. Curr. Biol. *22*, 1000–1004.

Lara, A.H., and Wallis, J.D. (2015). The role of prefrontal cortex in working memory: a mini review. Front. Syst. Neurosci. *9*, 173.

Leavitt, M.L., Mendoza-Halliday, D., and Martinez-Trujillo, J.C. (2017). Sustained activity encoding working memories: not fully distributed. Trends Neurosci. *40*, 328–346.

Lee, J.H., Whittington, M.A., and Kopell, N.J. (2013). Top-down beta rhythms support selective attention via interlaminar interaction: a model. PLoS Comput. Biol. *9*, e1003164.

Leszczynski, M., Fell, J., Jensen, O., and Axmacher, N. (2017). Alpha activity in the ventral and dorsal visual stream controls information flow during working memory. bioRxiv. https://doi.org/10.1101/180166.

Lisman, J.E. (1997). Bursts as a unit of neural information: making unreliable synapses reliable. Trends Neurosci. *20*, 38–43.

Lisman, J.E., and Idiart, M.A. (1995). Storage of 7 +/- 2 short-term memories in oscillatory subcycles. Science *267*, 1512–1515.

Luck, S.J., and Vogel, E.K. (1997). The capacity of visual working memory for features and conjunctions. Nature *390*, 279–281.

Lundqvist, M., Herman, P., and Lansner, A. (2011). Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. J. Cogn. Neurosci. *23*, 3008–3020.

Lundqvist, M., Herman, P., and Lansner, A. (2012). Variability of spike firing during θ-coupled replay of memories in a simulated attractor network. Brain Res. *1434*, 152–161.

Lundqvist, M., Rose, J., Herman, P., Brincat, S.L., Buschman, T.J., and Miller, E.K. (2016). Gamma and Beta Bursts Underlie Working Memory. Neuron *90*, 152–164.

Lundqvist, M., Herman, P., Warden, M.R., Brincat, S.L., and Miller, E.K. (2018a). Gamma and beta bursts during working memory readout suggest roles in its volitional control. Nat. Commun. *9*, 394.

Lundqvist, M., Herman, P., and Miller, E.K. (2018b). Working memory: delay activity, yes! persistent activity? maybe not. J. Neurosci. *38*, 7013–7019.

Maier, A., Adams, G.K., Aura, C., and Leopold, D.A. (2010). Distinct superficial and deep laminar domains of activity in the visual cortex during rest and stimulation. Front. Syst. Neurosci. *4*, 31.

Markov, N.T., Ercsey-Ravasz, M., Van Essen, D.C., Knoblauch, K., Toroczkai, Z., and Kennedy, H. (2013). Cortical high-density counterstream architectures. Science *342*, 1238406.

Markov, N.T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., Lamy, C., Misery, P., Giroud, P., Ullman, S., et al. (2014). Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. J. Comp. Neurol. *522*, 225–259.

Mejias, J.F., Murray, J.D., Kennedy, H., and Wang, X.-J. (2016). Feedforward and feedback frequency-dependent interactions in a large-scale laminar network of the primate cortex. Sci. Adv. *2*, e1601335.

Mendoza-Halliday, D., Torres, S., and Martinez-Trujillo, J.C. (2014). Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. Nat. Neurosci. *17*, 1255–1262.

Meyers, E.M., Freedman, D.J., Kreiman, G., Miller, E.K., and Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. J. Neurophysiol. *100*, 1407–1419.

Mi, Y., Katkov, M., and Tsodyks, M. (2017). Synaptic correlates of working memory capacity. Neuron *93*, 323–330.

Michalareas, G., Vezoli, J., van Pelt, S., Schoffelen, J.-M., Kennedy, H., and Fries, P. (2016). Alpha-beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical areas. Neuron *89*, 384–397.

Miller, E.K., and Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. *24*, 167–202.

Miller, E.K., and Wilson, M.A. (2008). All my circuits: using multiple electrodes to understand functioning neural networks. Neuron *60*, 483–488.

Miller, E.K., Erickson, C.A., and Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. J. Neurosci. *16*, 5154–5167.

Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic theory of working memory. Science *319*, 1543–1546.

Murray, J.D., Bernacchia, A., Freedman, D.J., Romo, R., Wallis, J.D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D., and Wang, X.J. (2014). A hierarchy of intrinsic timescales across primate cortex. Nat. Neurosci. *17*, 1661–1663.

Murray, J.D., Bernacchia, A., Roy, N.A., Constantinidis, C., Romo, R., and Wang, X.-J. (2017). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. Proc. Natl. Acad. Sci. USA *114*, 394–399.

Naud, R., and Sprekeler, H. (2018). Sparse bursts optimize information transmission in a multiplexed neural code. Proc. Natl. Acad. Sci. USA *115*, E6329–E6338.

Parnaudeau, S., O'Neill, P.-K., Bolkan, S.S., Ward, R.D., Abbas, A.I., Roth, B.L., Balsam, P.D., Gordon, J.A., and Kellendonk, C. (2013). Inhibition of mediodorsal thalamus disrupts thalamofrontal connectivity and cognition. Neuron *77*, 1151–1162.

Parthasarathy, A., Herikstad, R., Bong, J.H., Medina, F.S., Libedinsky, C., and Yen, S.-C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. Nat. Neurosci. *20*, 1770–1779.

Pasternak, T., and Greenlee, M.W. (2005). Working memory in primate sensory systems. Nat. Rev. Neurosci. *6*, 97–107.

Popov, T., Kastner, S., and Jensen, O. (2017). FEF-controlled alpha delay activity precedes stimulus-induced gamma-band activity in visual cortex. J. Neurosci. *37*, 4117–4127.

Richter, C.G., Thompson, W.H., Bosman, C.A., and Fries, P. (2017). Top-down beta enhances bottom-up gamma. J. Neurosci. *37*, 6698–6711.

Richter, C.G., Coppola, R., and Bressler, S.L. (2018). Top-down beta oscillatory signaling conveys behavioral context in early visual cortex. Sci. Rep. *8*, 6991.

Rigotti, M., Barak, O., Warden, M.R., Wang, X.-J., Daw, N.D., Miller, E.K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature *497*, 585–590.

Rockland, K.S., and Pandya, D.N. (1979). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. Brain Res. *179*, 3–20.

Roesch, M.R., and Olson, C.R. (2005). Neuronal activity dependent on anticipated and elapsed delay in macaque prefrontal cortex, frontal and supplementary eye fields, and premotor cortex. J. Neurophysiol. *94*, 1469–1497.

Romo, R., Brody, C.D., Hernández, A., and Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. Nature *399*, 470–473.

Rose, N.S., LaRocque, J.J., Riggall, A.C., Gosseries, O., Starrett, M.J., Meyering, E.E., and Postle, B.R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. Science *354*, 1136–1139.

Roux, F., Wibral, M., Mohr, H.M., Singer, W., and Uhlhaas, P.J. (2012). Gamma-band activity in human prefrontal cortex codes for the number of relevant items maintained in working memory. J. Neurosci. *32*, 12411–12420.

Salazar, R.F., Dotson, N.M., Bressler, S.L., and Gray, C.M. (2012). Content-specific fronto-parietal synchronization during visual working memory. Science *338*, 1097–1100.

Sandberg, A., Tegnér, J., and Lansner, A. (2003). A working memory model based on fast Hebbian learning. Network *14*, 789–802.

Schmitt, L.I., Wimmer, R.D., Nakajima, M., Happ, M., Mofakham, S., and Halassa, M.M. (2017). Thalamic amplification of cortical connectivity sustains attentional control. Nature *545*, 219–223.

Schroeder, C.E., Wilson, D.A., Radman, T., Scharfman, H., and Lakatos, P. (2010). Dynamics of active sensing and perceptual selection. Curr. Opin. Neurobiol. *20*, 172–176.

Shafi, M., Zhou, Y., Quintana, J., Chow, C., Fuster, J., and Bodner, M. (2007). Variability in neuronal activity in primate cortex during working memory tasks. Neuroscience *146*, 1082–1108.

Sherman, M.A., Lee, S., Law, R., Haegens, S., Thorn, C.A., Hämäläinen, M.S., Moore, C.I., and Jones, S.R. (2016). Neural mechanisms of transient neocortical beta rhythms: converging evidence from humans, computational modeling, monkeys, and mice. Proc. Natl. Acad. Sci. USA *113*, E4885–E4894.

Siegel, M., Warden, M.R., and Miller, E.K. (2009). Phase-dependent neuronal coding of objects in short-term memory. Proc. Natl. Acad. Sci. USA *106*, 21341–21346.

Siegel, M., Buschman, T.J., and Miller, E.K. (2015). Cortical information flow during flexible sensorimotor decisions. Science *348*, 1352–1355.

Sigala, N. (2009). Visual working memory and delay activity in highly selective neurons in the inferior temporal cortex. Front. Syst. Neurosci. *3*, 11.

Smith, M.A., Jia, X., Zandvakili, A., and Kohn, A. (2013). Laminar dependence of neuronal correlations in visual cortex. J. Neurophysiol. *109*, 940–947.

Spaak, E., Bonnefond, M., Maier, A., Leopold, D.A., and Jensen, O. (2012). Layer-specific entrainment of γ-band neural activity by the α rhythm in monkey visual cortex. Curr. Biol. *22*, 2313–2318.

Spaak, E., Watanabe, K., Funahashi, S., and Stokes, M.G. (2017). Stable and dynamic coding for working memory in primate prefrontal cortex. J. Neurosci. *37*, 6503–6516.

Spitzer, B., and Haegens, S. (2017). Beyond the status quo: a role for beta oscillations in endogenous content (re)activation. eNeuro *4*. Published online August 2, 2017. https://doi.org/10.1523/ENEURO.0170-17.2017.

Sprague, T.C., Ester, E.F., and Serences, J.T. (2016). Restoring latent visual working memory representations in human cortex. Neuron *91*, 694–707.

Stanley, D.A., Roy, J.E., Aoi, M.C., Kopell, N.J., and Miller, E.K. (2018). Low-beta oscillations turn up the gain during category judgments. Cereb. Cortex *28*, 116–130.

Stokes, M.G. (2015). 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn. Sci. *19*, 394–405.

Stokes, M., and Spaak, E. (2016). The importance of single-trial analyses in cognitive neuroscience. Trends Cogn. Sci. *20*, 483–486.

Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., and Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. Neuron *78*, 364–375.

Uhlhaas, P.J., and Singer, W. (2010). Abnormal neural oscillations and synchrony in schizophrenia. Nat. Rev. Neurosci. *11*, 100–113.

VanRullen, R. (2016). Perceptual cycles. Trends Cogn. Sci. *20*, 723–735.

Vogel, E.K., and Machizawa, M.G. (2004). Neural activity predicts individual differences in visual working memory capacity. Nature *428*, 748–751.

Voytek, B., Kayser, A.S., Badre, D., Fegen, D., Chang, E.F., Crone, N.E., Parvizi, J., Knight, R.T., and D'Esposito, M. (2015). Oscillatory dynamics coordinating human frontal networks in support of goal maintenance. Nat. Neurosci. *18*, 1318–1324.

Wang, X.-J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. Physiol. Rev. *90*, 1195–1268.

Wang, X.-J., and Yang, G.R. (2018). A disinhibitory circuit motif and flexible information routing in the brain. Curr. Opin. Neurobiol. *49*, 75–83.

Wang, Y., Markram, H., Goodman, P.H., Berger, T.K., Ma, J., and Goldman-Rakic, P.S. (2006). Heterogeneity in the pyramidal network of the medial prefrontal cortex. Nat. Neurosci. *9*, 534–542.

Warden, M.R., and Miller, E.K. (2010). Task-dependent changes in short-term memory in the prefrontal cortex. J. Neurosci. *30*, 15801–15810.

Wasmuht, D.F., Spaak, E., Buschman, T.J., Miller, E.K., and Stokes, M.G. (2018). Intrinsic neuronal dynamics predict distinct functional roles during working memory. Nat. Commun. *9*, 3499.

Watanabe, K., and Funahashi, S. (2014). Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. Nat. Neurosci. *17*, 601–611.

Watanabe, Y., and Funahashi, S. (2004). Neuronal activity throughout the primate mediodorsal nucleus of the thalamus during oculomotor delayed-responses. I. Cue-, delay-, and response-period activity. J. Neurophysiol. *92*, 1738–1755.

Watrous, A.J., Fell, J., Ekstrom, A.D., and Axmacher, N. (2015). More than spikes: common oscillatory mechanisms for content specific neural representations during perception and memory. Curr. Opin. Neurobiol. *31*, 33–39.

Whittington, M.A., Traub, R.D., Kopell, N., Ermentrout, B., and Buhl, E.H. (2000). Inhibition-based rhythms: experimental and mathematical observations on network dynamics. Int. J. Psychophysiol. *38*, 315–336.

Wolff, M.J., Jochim, J., Akyürek, E.G., and Stokes, M.G. (2017). Dynamic hidden states underlying working-memory-guided behavior. Nat. Neurosci. *20*, 864–871.

Woloszyn, L., and Sheinberg, D.L. (2009). Neural dynamics in inferior temporal cortex during a visual working memory task. J. Neurosci. *29*, 5494–5507.

Wutz, A., Loonis, R., Roy, J.E., Donoghue, J.A., and Miller, E.K. (2018). Different levels of category abstraction by different dynamics in different prefrontal areas. Neuron *97*, 716–726.e8.

Xing, D., Yeh, C.-I., Burns, S., and Shapley, R.M. (2012). Laminar analysis of visually evoked activity in the primary visual cortex. Proc. Natl. Acad. Sci. USA *109*, 13871–13876.

Zaldivar, D., Goense, J., Lowe, S.C., Logothetis, N.K., and Panzeri, S. (2018). Dopamine is signaled by mid-frequency oscillations and boosts output layers visual information in visual cortex. Curr. Biol. *28*, 224–235.e5.

Zhang, Y., Chen, Y., Bressler, S.L., and Ding, M. (2008). Response preparation and inhibition: the role of the cortical sensorimotor beta rhythm. Neuroscience *156*, 238–246.